ABSTRACT

Described herein is a technology for recognizing the content of text documents. The technology determines one or more hash values for the content of a text document. Alternatively, the technology may generate a "sifted text" version of a document. In one implementation described herein, document recognition is used to determine whether the content of one document is copied (i.e., plagiarized) from another document. This is done by comparing hash values of documents (or alternatively their sifted text). In another implementation described herein, document recognition is used to categorize the content of a document so that it may be grouped with other documents in the same category. This abstract itself is not intended to limit the scope of this patent. The scope of the present invention is pointed out in the appending claims.